N92-22327

PERFORMANCE ASSESSMENT IN COMPLEX INDIVIDUAL AND TEAM TASKS

Douglas R. Eddy, Ph.D.
NTI, Incorporated
P.O. Box 35482
Brooks AFB Texas

ABSTRACT

This paper describes an eclectic, performance-based approach to assessing cognitive performance from multiple perspectives. The experience gained from assessing the effects of antihistamines and scenario difficulty on C^2 decision-making performance in Airborne Warning and Control Systems (AWACS) weapons director (WD) teams can serve as a model for realistic simulations in space operations. Emphasis is placed on the flexibility of measurement, hierarchical organization of measurement levels, data collection from multiple perspectives, and the difficulty of managing large amounts of data.

INTRODUCTION

Astronomers in the late 1700s recorded a star's transit by using a metronome to determine the moment a star touched the cross hairs of the telescope. The chief astronomer at Greenwich Observatory noticed that his assistant's times were consistently one second slower than his own. This was an early realization that the observer played a significant role in acquiring data and that even simple perceptual observations were susceptible to bias and individual abilities. Today we have sophisticated instruments to record much of the data of interest to science. However, in complex tasks where decisions must be based on human judgement or on the consensus of a team, the roles of the integrator of information and the decision-maker are still important and still susceptible to bias.

In the implementation of large projects such as building and maintaining a space station, building and maintaining a moon colony, or traveling to Mars, various designers need to know how our astronauts will handle the work. Engineers want to design consoles and workstations so operators can perform their tasks efficiently and without errors. Trainers want to provide timely and objective feedback to operators. Mission planners want to design work/rest cycles that maximize productivity while minimizing error and waste. Social planners want to provide work environments that facilitate team interaction and cooperation while minimizing the disruption of violations of personal space and privacy. The only way designers can have confidence about how our astronauts will perform on complex tasks far, far from home is by assessing performance in early design studies. Further, to maximize the use of equipment, facilities, subject time, and to obtain the most integrated data possible, complex, realistic, ground-based studies involving integrated payloads will be required. These future studies can benefit from the approach used in the Crew Performance Branch at Brooks Air Force Base, Texas, to assess antihistamine effects on complex task performance in WD teams and from lessons learned in the study. Since this paper is designed to communicate methods and approaches to understanding complex team tasks, emphasis is placed on experimental design issues with only sample results presented.

Although the primary goals of the study were to evaluate the effects of Seldane on complex performance, the researchers used the opportunity to gather data on several other issues from several perspectives. These included: the development of a methodology for assessing individual and team complex-task performance, the evaluation of sustained operations and fatigue, the assessment of cognitive workload through embedded tasks, the assessment of stress, the assessment of learning effects, the evaluation of tests for WD selection, and the prediction of complex task performance from cognitive skills tests.

THE WEAPONS DIRECTOR TASK

WDs in an air defense scenario must attend to a number of tasks. The wartime tasks include locating and identifying aircraft, maintaining track information on aircraft and targets, updating target information received from pilots, accepting aircraft hand-offs, performing a tactical controller function with appropriate level of control, providing target briefings to interceptors, performing a tanker controller function, providing recovery assistance, safe passage monitoring, briefing the senior director (SD) of any tracking or sensor data problems, and responding to alerts, alarms, and messages on the console. The success of the C^2 mission results directly from the WDs' successful accomplishment of their duties.

PERFORMANCE HIERARCHY

It is obvious that the performance of such a complex system including human operators is a result of numerous interacting internal and external factors. Because of these multiple determinants and numerous data perspectives, it was necessary to use a variety of measures to characterize the system and to diagnose the sources of observed variations in system performance. The interpretation of large metric sets is facilitated by an implicit underlying structure that weights the significance of each measure and relates it to the others.

After a review of the literature on objective measures of team performance (Eddy, 1989), the measurement aggregation problem was approached by devising a hierarchy of performance determinants that provides a classification framework for individual measures. Each level of the hierarchy contains groups of measures that jointly determine the measures available at the next level higher in the framework. For the command and control air defense scenario, four levels were chosen as shown in Figure 1. They are: Mission Effectiveness, System/team Performance, Individual Performance, and Performance Capability.

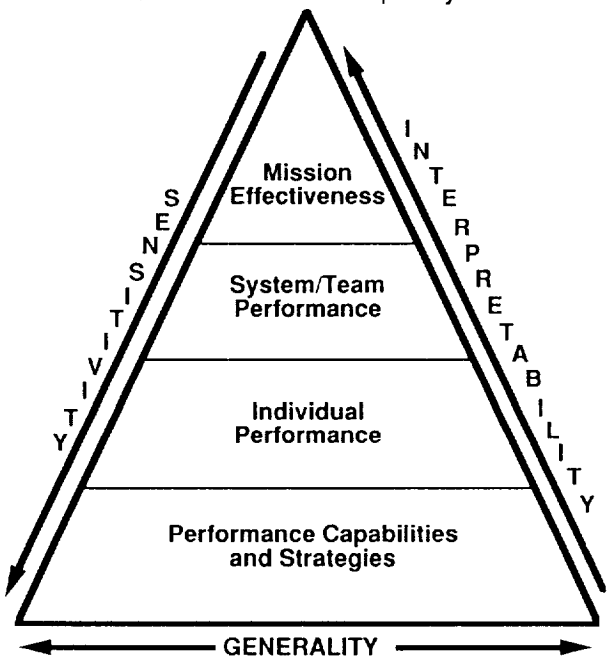

Figure 1. Performance measurement hierarchy.

The highest level of the hierarchy contains indices of Mission Effectiveness. These measures are derived directly from the specific objectives of the mission assigned to the system. An example is the protection of a specific sector of air and ground space from infiltration by enemy aircraft (protection of assets). Measures that flow from this objective and that assess performance in terms of mission effectiveness include the number of enemy infiltrations, the amount of fuel and weapons expended, and the ratio of enemy lost to friendly assets lost.

The second level of the hierarchy, System Performance, contains those groups of measures that reflect factors immediately affecting mission effectiveness. Such measures of System Performance reflect the degree to which the combined man-machine system has accomplished those tasks required to meet mission objectives. These measures do not reflect the individual contributions of different human behaviors or various hardware and software component performances. Instead, they are more global indices of the degree to which the total system successfully accomplished the tasks essential to mission success. For example, the weapons director/workstation system is required to meet its mission objectives essentially by accomplishing an air traffic control task aimed at directing interceptor aircraft to defeat threat aircraft. This air traffic control task decomposes into a number of essential subtasks such as pairing of interceptors with targets and providing target data to interceptors. A performance measure of the latter is the average accuracy and speed of data transfer to interceptor pilots.

The third level of the hierarchy is comprised of specific groups of measures that assess the individual contributions of human components to overall system performance. Measures included in the Human Performance level of the hierarchy are designed to reflect the quality of the individual behaviors required of the WD expressed primarily in terms of latencies and errors. These measures are derived by examining the system functions required to meet mission objectives in order to identify the specific contributions of the operator. For example, the system performance requirement to pair targets with interceptors requires the weapons director to identify a target's location on the workstation display and communicate this information to an interceptor aircraft via radio. The quality of the operator's performance in achieving this objective might be measured by evaluating the time needed to complete the full sequence of required behaviors and by assessing the accuracy of each manual and verbal response.

The final level of the proposed hierarchy contains measures that assess factors directly affecting the individual performance capacities of primary system components. For the human operator, measures of Performance Capability are composed of a large group of potential human state and ability measures that combine to determine overt performance. These measures include indices of workload or reserve processing capacity; fatigue; arousal level; experience level; and individual perceptual, cognitive, and motor abilities. This level also includes personality traits and predisposition to interact with teammates in specific ways that may or may not be adaptive under stress.

The multi-level classification of performance measures proposed above has the advantage of placing the measures into logical subordinate and superordinate groups indicating the predictive relationships among them. In addition, measures at each of the levels differ

in their sensitivity, generalizability, and practical interpretability.

REALISTIC SCENARIOS WITH EXPERIMENTAL CONTROL

Seldane, an antihistamine used in the study, does not cross the blood brain barrier and therefore may not affect performance. This puts the researcher in the unenviable position of trying to prove the null hypothesis. As a result it was necessary to demonstrate the sensitivity of the performance measures to some degrading treatment. Another antihistamine, Benadryl, was chosen for this purpose. (Merrell Dow Pharmaceuticals, Inc., provided the drugs for the study.) A placebo control brought the number of groups to three. Because it was possible to randomly assign experienced teams to one group and four inexperienced teams to another, we decided to collect data from each team under both difficulty conditions and under one antihistamine and placebo. Further, we wanted to collect two days of data under the drug conditions. Meeting these constraints required the development of six equivalent scenarios, but we had to prevent the subjects from anticipating the events in each succeeding scenario.

For the first scenario, we defined all the tracks, enemy flight paths, and events. We made lists of events and used a subject matter expert to indicate the impact on WD behavior of each. To prevent subjects from anticipating scenario events, in scenarios two through six we rotated the original so the enemy would appear from a different compass heading. Equivalent events were then spread across each scenario at the same points in time. We also changed land masses using different geographic locations, we created six unique prebriefings containing different political situations, countries, airbases, squadrons, call signs, and numbers. Debriefings at the end of all testing did not reveal that subjects believed any of the scenarios to be similar.

Scenario difficulty was manipulated in several ways. Enemy aircraft flew at varying altitudes and some took zigzag paths. The fog of war was increased by additional distractor events. Three scenarios were created for low difficulty and three for high difficulty.

METHODS

The 552d Air Wing assigned twelve teams of three WDs (male and female), who previously volunteered, to Brooks AFB to spend their work week in support of this study. The teams were randomly assigned to one of three drug treatment conditions and one of two scenario difficulty orders, either low-high or high-low.

The WDs arrived at Brooks AFB on either Saturday or Sunday evening for a preliminary briefing. Training took place on Monday for approximately eight hours. Teams received training on the AWACS-Performance Assessment Battery (AWACS-PAB), six simple computerized tests and two complex tests, over approximately four hours. The two complex tests were taken from the Complex Cognitive Assessment Battery that consists of nine tests.

The six tests were taken from Unified Triservice Cognitive Performance Assessment Battery (UTC-PAB) with over 25 tests. Further information on these tests can be obtained in Perez, Masline, Ramsey, and Urban (1987) and Hartel (1988). They also completed a three-hour C^3 training scenario to familiarize them with the simulated AWACS crewstations and scenarios. Subjects ingested one Benadryl and one Seldane placebo at 2230 or prior to going to sleep.

Starting on Tuesday, teams were then tested in two 3½ hour scenarios each day for three days. Each group ingested only placebos during the testing schedule for Tuesday. A randomly assigned team ingested the recommended therapeutic dose of either Benadryl, Seldane, or a lactose placebo starting on Tuesday evening. Total antihistamine/placebo ingestion for each group across two days consisted of either eight 25mg Benadryl, four 60mg Seldane, or all placebo preparations.

One SD was used for all teams. After the prebriefing, his interaction with the team was to give direction only when required, but to keep the team from straying outside the performance measurement envelope. Other details of the facilities, equipment, scenario development and time schedules may be found in Schiflett, Strome, Eddy, and Dalrymple (1990).

Because the cognitive performance of the weapons director teams can be interpreted for a variety of questions, several subject trait, experience, and state measures were recorded. These included: a biographical sketch, a WD experience form, personality scales for potential use in developing WD selection tools, and surveys of their current state (symptoms, sleepiness, fatigue, etc.). The scales included the Rotter Scale, which assesses the locus of control generally perceived by a person in causing changes to take place in one's life; the Personal Characteristics Inventory (PCI), which assesses attitudes and leadership qualities; the Life Style Questionnaire, which predicts a subject's performance under stress; the Least Preferred Co-worker Scale (LPC), which may identify a WD's leadership style; the Jenkins Activity Scale, which assesses a WD's personality characteristics of decision-making; and the FIRO-B, which measures a subject's attitudes with regard to sociability and social interaction.

WD ratings were also obtained on the USAFSAM Fatigue Scale, which allows the subject to describe how he/she feels at that time; an Operational Impact Survey, which allows a subject to rate how well he/she felt the team completed its mission and how well each subject felt he/she completed his/her part of the mission; a Scenario Evaluation form, which allowed each WD to order the simulations with respect to difficulty; and the Subject Workload Assessment Technique (SWAT), which allowed each subject, at the

end of each simulation, to evaluate the workload of the scenario along SWAT's three dimensions: time load, mental effort, and psychological stress. The WDs kept logs similar to those kept during a standard mission. They recorded aircraft call signs, type aircraft, target numbers paired against, check-in time, weapons states on the aircraft at RTB, results, and other information.

In addition to the outcome measures of how well a team or individual is performing in a simulated air defense scenario, one would like to understand the underlying processes that contribute to those outcomes. Embedded tasks were used to measure *reserve capacity*, team coordination, and *situational awareness* (SA). These are tasks natural to the air defense scenario, but low priority. These tasks were delivered auditorially by voice queries articulated by the Votan speech synthesizer or by the SD.

The embedded measures for reserve capacity are: 1. whether or not a response is given, 2. accuracy of the response, and 3. latency of the response. The independent variables that may determine the WDs' workload level are: the number of flights currently under the WD's control, the level of control of each flight, the ADWL, and the number and type of additional tasks currently being worked by the WD. A typical SD query for reserve capacity might be "What state armament/fuel on the aircraft under your control?" Low difficulty should result in quick, accurate responses from the WD. High difficulty should result in ignored requests, partial information, and long response times.

Individual members of a WD team can work independently of each other. However, since the enemy is directing the attack in an air defense scenario, the battle does not always unfold the way it is planned in a mission prebriefing. As a result, each WD's responsibilities change throughout the mission. These changes should be adaptive and result from insight and leadership. Further, the adaptations require cooperation and coordination among the team members. WD responses involve passing and confirming information to each other and accepting responsibility for incoming requests when time is available. Embedded measures for team coordination include: 1. Whether or not the information is passed to the other WDs, 2. Accuracy of the response, and 3. Latency of the response. An event designed to elicit a team coordination response might be an ADWL announcement from ground control.

To effectively deal with events in an air defense scenario, a WD must maintain an accurate representation of the battle. This representation (both internal memory and external notes) defines the WD's awareness of the current situation. If the representation is in error, the WD may commit to kill rather than identify an unknown target. Therefore, throughout the scenario the WD's awareness was probed to determine if he/she has the correct ADWL, has kept track of airbase openings/closings, and tracked SAM sites going hot/cold. The embedded measures for situational awareness are the same as for difficulty. An event designed to elicit a response would be for the SD to tell WD1 to kill track 0304. The WD should question this command during peace time since the SD had no authority to issue the order.

RESULTS

Early results for Mission and System/team performance levels and Performance Capability level have shown differential sensitivity to drug effects. At the Mission Effectiveness and System/team levels, 33 dependent measures, were amenable to statistical analysis. Of these measures, 6 showed a scenario difficulty effect, 4 showed a learning effect (days), and 8 showed a day by difficulty interaction. Table I shows the enemy penetrations, "get throughs," by day and difficulty. Although this variable did not achieve statistically significant results, it dramatically shows the impact of scenario difficulty and of performance improvement across days.

Table I. Enemy penetrations by day and scenario difficulty for all teams.

Condition	Penetrations
Day 2	
High Difficulty	22
Low Difficulty	5
Day 3	
High Difficulty	13
Low Difficulty	3
Day 4	
High Difficulty	6
Low Difficulty	6

Figure 2 shows the effect of scenario difficulty on the loss ratio of enemy/friendly aircraft. Loss ratios remained the same across days while ratios improved across days under low difficulty. In no case did performance under either antihistamine differ from the placebo group. These performance results for scenario difficulty were supported with WD ratings using the Subjective Workload Assessment Technique (SWAT). Generally scenarios designed for high difficulty resulted in higher workload ratings than those designed for low difficulty. A full description of the results can be found in Eddy, Dalrymple, and Schiflett (in preparation).

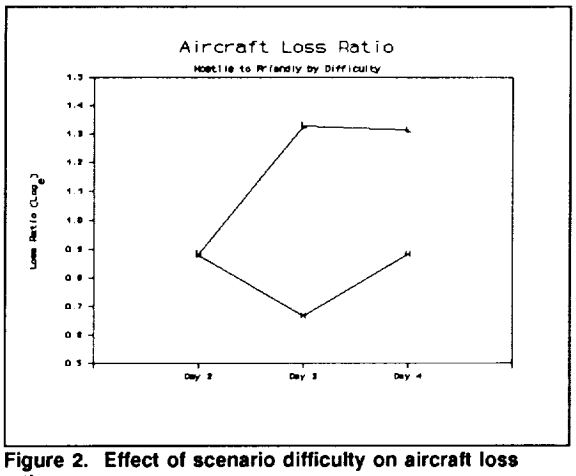

Figure 2. Effect of scenario difficulty on aircraft loss ratio.

By rearranging the data, time-of-day effects could be analyzed. Interestingly, performance on the morning simulations did not differ from that in the evenings, with difficulty balanced, even though subjective fatigue measures were higher during the evening simulation.

Benadryl degraded performance on cognitive skills and abilities as measured by the AWACS-PAB, especially on the first day of Benadryl administration, day 3 (Nesthus, Schiflett, Eddy, Whitmore, in preparation). Six of the tests, eight of the dependent measures, showed either a significant drug and/or drug-by-day effect. For example, Figure 3 shows an increase in errors in the Benadryl group on the Dichotic Listening test. Figure 4 shows that the Benadryl group found fewer word solutions on medicated days than the other groups who improved their performance on their treatment days. In addition, the Benadryl subjects subjective assessment of fatigue was greater on day 3. Seldane had no effect on performance as measured by these tests.

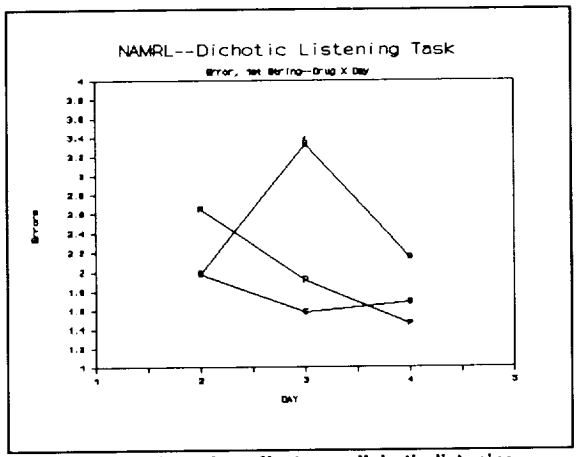

Figure 3. Antihistamine effects on dichotic listening.

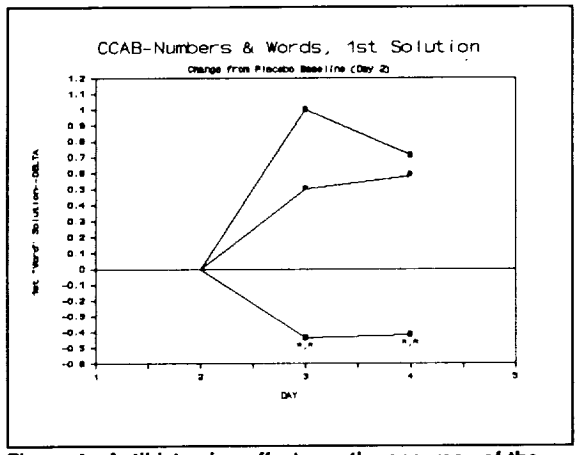

Figure 4. Antihistamine effects on the accuracy of the first solution in the Numbers and Words test.

FUTURE DIRECTIONS FOR THE AWACS DATA

The next step in data analysis involves developing rules for assigning individual WD responsibility within each scenario. These rules or definitions of areas of responsibility follow from the WDs training and practice. Once developed, each individual's role in "winning the war" can be assessed. This will include how well a WD controls his or her own area of responsibility (AOR), how he or she assists others, and how he or she requests assistance from the WD team. Through this approach the team's performance can be understood as a combination of individual efforts that either support or block the attainment of team goals. After the outcome measures of individual performance are obtained, process measures on the WD tasks and subtasks that produce the outcomes will be assessed. These measures will assess how well the individuals and teams accomplish such tasks as committing interceptors to targets, passing information to pilots, conducting intercepts, maintaining coverage of CAP points, maintaining situational awareness, etc.

MEASUREMENT PROBLEMS IN COMPLEX TASKS

To answer the question of why outcome measures of a system or team are degraded or improved, one must plunge into task analysis and modelling. Once good models of WD and team processes are established, objective measures of those processes can be evaluated against criterion individual and team outcome measures. Currently programmers and researchers in the AESOP laboratory are reviewing individual outcome measures and decomposing WD tasks and processes.

Several problems arise in attempting to objectively measure complex processes. For example, the beginning and ending of a task or process may not have well defined criteria or may cross media boundaries. We have often found that we can start with something concrete in the process, such as a switch action, and then work forward and backward for the start and end of the process. Sometimes this involves locating the switch action in the data file, obtaining the time stamp, using the time stamp to search through a file of transcribed utterances, and finally locating the initiating and/or ending event. This is a labor intensive process, but has the potential of being automated in the future with a text parser.

Another problem that arises involves simultaneous or overlapping tasks. Identifying when this happens and analyzing the single and dual tasks separately is one solution. If the same tasks overlap frequently and one task has a low priority, it may be possible to use the low priority task as an embedded secondary task to assess reserve capacity. Important tasks and processes that occur infrequently can provide highly variable latencies. If these tasks and processes have similar effects on the WD's behavior it may be possible to collapse the latencies of several treating them as a group. Often one task will interrupt another. This is an opportunity to verify the subject's prioritization of these tasks and if enough data exits, a confusability matrix can be generated.

LESSONS LEARNED FOR SPACE OPERATIONS

Because of the needs mentioned at the beginning of the paper, greater emphasis will be placed on understanding the effects of individual components on the performance of a complex system. This in turn calls out for the conduct of experiments with integrated payloads and performance measures to answer questions from multiple perspectives. As researchers, we must meet these needs by developing methods to assess performance in complex tasks. Our research on AWACS WDs has demonstrated that errors, failures, breakdowns in procedures, and systems may not show up unless the system is stressed. Researchers in space operations must continually search for system stressors that are realistic and appropriate to test a system's performance and its components. In this regard, statistical designs with repeated measures will be necessary to reduce variability, thereby requiring sophisticated ways of preventing subjects from anticipating events in repeated scenarios.

REFERENCES

Eddy, D. E. (1989). Selected team performance measures in a C^3 Environment--An Annotated Bibliography. Technical Report USAFSAM-TR-87-25, USAF School of Aerospace Medicine, Brooks AFB, Texas.

Hartel, C. (1988). Expanded complex cognitive assessment battery (CCAB): Test descriptions. AAC-UM-33221, Systems Research Laboratories, U.S. Army Research Institute, VA: Alexandria.

Nesthus, T. E., Schiflett, S. G., Eddy, D. R., and Whitmore, J. N. (in press). Comparative effects of antihistamines on aircrew performance of simple and complex tasks under sustained operations.

Perez, W. A., Masline, P. J., Ramsey, E. G., and Urban, K. E. (1987). Unified tri-services cognitive performance assessment battery: Review and methodology. Technical Report AAMRL-TR-87-007, Armstrong Aerospace Medical Research Laboratory, Wright-Patterson AFB, Ohio.

Schiflett, S. G., Strome, D. S., Eddy, D. R., and Dalrymple, M. A. (1990). Aircrew evaluation sustained operations performance (AESOP): A Triservice Facility for Technology Transition. Technical Paper USAFSAM-TP-90-26, USAF School of Aerospace Medicine, Brooks AFB, Texas.